ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Research Article

Identification of Antibiotic Resistance Proteins via MiCId's Augmented Workflow. A Mass Spectrometry-Based Proteomics Approach

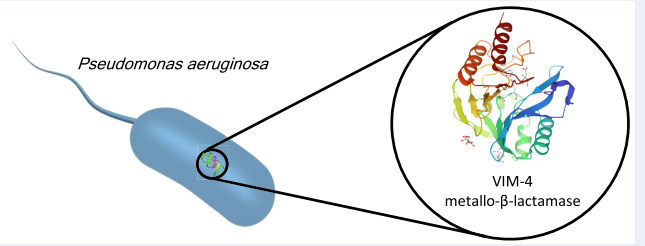
Gelio Alves, Aleksey Ogurtsov, Roger Karlsson, Daniel Jaén-Luchoro, Beatriz Piñeiro-Iglesias, Francisco Salvà-Serra, Björn Andersson, Edward R. B. Moore, and Yi-Kuo Yu*

Cite This: *J. Am. Soc. Mass Spectrom.* 2022, 33, 917−931

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Fast and accurate identifications of pathogenic bacteria along with their associated antibiotic resistance proteins are of paramount importance for patient treatments and public health. To meet this goal from the mass spectrometry aspect, we have augmented the previously published *Mi*croorganism Classification and *Id*entification (MiCId) workflow for this capability. To evaluate the performance of this augmented workflow, we have used MS/MS datafiles from samples of 10 antibiotic resistance bacterial strains belonging to three different species: *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*. The evaluation shows that MiCId's workflow has a sensitivity value around 85% (with a lower bound at about 72%) and a precision greater than 95% in identifying antibiotic resistance proteins. In addition to having high sensitivity and precision, MiCId's workflow is fast and portable, making it a valuable tool for rapid identifications of bacteria as well as detection of their antibiotic resistance proteins. It performs microorganismal identifications, protein identifications, sample biomass estimates, and antibiotic resistance protein identifications in 6−17 min per MS/MS sample using computing resources that are available in most desktop and laptop computers. We have also demonstrated other use of MiCId's workflow. Using MS/MS data sets from samples of two bacterial clonal isolates, one being antibiotic-sensitive while the other being multidrug-resistant, we applied MiCId's workflow to investigate possible mechanisms of antibiotic resistance in these pathogenic bacteria; the results showed that MiCId's conclusions agree with the published study. The new version of MiCId (v.07.01.2021) is freely available for download at https://www.ncbi.nlm.nih.gov/CBBresearch/Yu/downloads.html.



KEYWORDS: *identification of antibiotic resistance proteins, microorganism identification/classification workflow, mass spectrometry*

1. INTRODUCTION

Fast and accurate identification of pathogenic bacteria along with the identification of antibiotic resistance (AR) proteins is of paramount importance for patient treatments and public health.[1−5] Once the pathogenic bacteria causing the infections are identified swiftly along with their AR proteins (if present), proper treatment can be administered, which can increase patients' survival rate and minimize improper use of antibiotics.[6,7]

Currently, molecular methods such as next-generation sequencing (NGS) and mass spectrometry (MS) are used and are being developed to speed up identifications of pathogenic bacteria.[8−25] While several computational workflows/pipelines for analyzing NGS data have been developed to identify pathogenic bacteria and AR genes,[26−28] a mass spectrometry workflow with this capability is still lacking.[24] This has motivated us to augment the workflow of our pathogen identification tool, *Mi*croorganism Classification and *Id*entification (MiCId),[21,29,30] to enable the identification of AR proteins using MS data from a high-performance liquid chromatography system coupled to a high-resolution tandem MS (HPLC−MS/MS). Another motivation for augmenting MiCId's workflow is that, even though NGS workflows can provide information about the presence of AR genes, they do not provide information about protein expression, which is extremely important for treating infections and for understanding the mechanism of antibiotic resistance in bacteria.[31−34]

For a summary of some of the existing workflows employed for the identification of bacteria using HPLC−MS/MS experiments, we refer readers to previous publications.[24,29,30] Overall, there has been significant progress made in the

Received: November 17, 2021
Revised: February 17, 2022
Accepted: February 18, 2022
Published: May 2, 2022

identification of bacteria using HPLC−MS/MS experiments, although there is plenty of room for improvement in sample preparation protocols and data analysis workflows.[35−37] Developers of HPLC−MS/MS data analysis workflows often use the sensitivity (true positive rate) and specificity (true negative rate) as the only criteria to assess the usability of the developed workflow. Although sensitivity and specificity are acceptable criteria to measure the performance of a workflow, these criteria alone are not enough to justify the usability of a workflow. For example, an important criterion that is often not mentioned in performance evaluations is the execution time. Identification of bacteria is a computationally demanding task for a workflow, as it has to query tens of thousands of MS/MS spectra in a microorganismal database containing thousands to tens of thousands of bacteria. In order to scale with the number of HPLC−MS/MS experiments, a workflow with appropriate amount of computer resources must have execution time less than the time it takes to conduct the HPLC−MS/MS experiment, which is approximately 1−2 h. This remained an unattainable goal for most workflows.[38] Other criteria to consider include whether or not a workflow provides for identified bacterial biomass estimation,[39,30] protein identification[40,29] with protein quantification,[41,42] and AR protein identification.[24] Data on the relative biomasses of identified bacteria identified are essential for studying microbial communities[39] and are valuable when determining treatment options for patients suffering from coinfections.[43−45] Knowledge of proteins and protein expression levels are essential for analyzing gene expression and function[46,47] and for investigating possible mechanisms of antibiotic resistance in bacteria.[31−34] Information about AR proteins is crucial for proper treatments for AR-resistant bacterial infections.[6,7] The criteria above cover most of the data analysis features needed for a workflow to be useful. In order to ensure a workflow to be user-friendly, intuitive, and customizable, we propose additional criteria. A useful workflow should: (1) automate and customize microorganismal protein sequences for download and database construction; (2) automate and customize AR protein sequences for download and database construction; (3) be computationally efficient and scalable to handle large microorganismal databases, large numbers of MS/MS spectra, and large number of MS/MS experiments; (4) be available to execute in different computer operating systems; (5) offer a user-friendly graphical interface. Meeting these latter criteria allows a workflow to eliminate elaborate intermediate steps and reaching a broader group of users in addition to experts in the field.

In previous studies, we have demonstrated that MiCId's workflow meets most of the criteria listed above.[21,29,30,48] We have shown that MiCId's workflow:

- Offers automated microorganismal database construction by automatically downloading from the NCBI database protein sequences of organisms specified by the user.
- Offers customized microorganismal database construction using a list of protein sequence Fasta files of organisms specified by the user that are stored in the local computer.
- Is able to identify bacteria in samples containing single and multiple bacteria with high sensitivity and high specificity by computing, for each identified taxon, an *E*-value which can be used to control the proportion of false discoveries (PFD) without the need of a decoy database.[21,29] When a list of candidate taxa are ranked by a quality score S, the *E*-value $E(S \geq S_0)$ is defined as the expected number of random taxa with scores the same as or better than S_0.
- Is able to estimate taxonomic biomass by computing a quantity called the *prior* using a modified expectation-maximization (EM) method. The *prior* is defined as the probability for a taxon to emit any evidence peptide and can be regarded as the taxon's relative protein biomass within the sample analyzed.[30]
- Provides protein identifications via combining peptides' *E*-values, using theoretically derived mathematical formulas.[40,49]
- Is computationally efficient and scalable, taking 6−17 min to process tens of thousands of MS/MS spectra in a large database, using resources available in most desktop/laptop computers.
- Is a self-contained workflow available with a friendly graphical user interface (GUI) with many features available for data analysis and visualization.

However, the previous versions of MiCId's workflow do not provide protein quantification or AR protein identification and are only available for the Linux operating system. In this study, we have augmented the MiCId's workflow to meet the criterion for the identification of AR proteins, and we intend to address the other two unmet criteria in the near future. MiCId's workflow can, however, be used in the Windows operating system via a virtual machine. Details of how to run MiCId's workflow in the Windows operating system are described in MiCId's user manual.

The AR protein identification task for an MS/MS workflow can be formulated as follows. First, using data from an MS/MS experiment, a workflow needs to identify the species/strains present in the biological sample. Second, it needs to construct, on the fly, a target protein database to be used for AR protein identifications. Even if a workflow has high sensitivity and high specificity for the identification of microorganisms and proteins, a remaining difficulty to be dealt with in identification of AR proteins is deciding what protein sequences to include in the target protein database. In principle, the ideal target protein database to use would include all of the protein sequences obtained directly from the strains present in the biological sample and with AR proteins unambiguously annotated. However, such a database is unobtainable from an MS/MS based proteomics approach, even if strain level identification is attained. It is standard practice for workflows to use databases such as those hosted by the National Center for Biotechnology Information (NCBI) to obtain protein sequences for as-yet-to-be-identified strains to build a target protein database. A target protein database constructed by using this procedure is an approximation to the ideal target protein database because the strains present in the biological sample could have gained new proteins via horizontal gene transfer and mutations through rapid multiplication and environmental pressure.[50,51] To mitigate this issue, MiCId constructs on the fly a target protein database made of proteins from the reference/representative proteomes of confidently identified species and AR proteins from a high-quality AR database.[27,52,53] This strategy is employed because the proteomes of reference/representative strains are proteome assemblies of higher quality; hence, they are to be used as anchors for the analysis of closely related proteomes within the same taxonomic group.[54] By including a

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

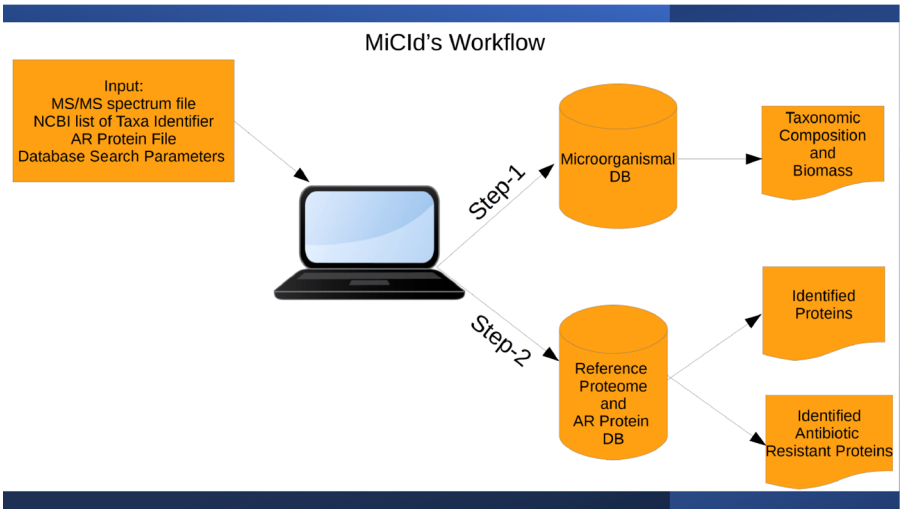


Figure 1. MiCId's workflow overview. To execute MiCId, users must provide the following input: a list of taxonomic identifiers taken from the NCBI, an experimental datafile (containing MS/MS spectra) of a microorganism sample, an antibiotic resistance (AR) protein file, and the parameters for database search. The list of taxonomic identifiers is used by MiCId to download from the NCBI the Fasta files of the protein sequences for all the taxa specified along with their taxonomic information. The downloaded protein Fasta files and the taxonomic file are used to create the microorganismal database. In step 1, the MS/MS spectra are queried in the microorganismal database in order to determine the taxonomic composition (via an iterative approach that propagate only taxa identified at one level to identifications at the next level) and the relative biomasses of microorganisms in the sample.[29,30] In step 2, the newly augmented step, MiCId generates a protein database that includes protein sequences from reference/representative strain of species identified with *E*-value ≤ 0.01 and *prior* ≥ 0.01 and from the user-specified AR protein file. The MS/MS spectra are then used to query this database to perform protein identifications, AR proteins included.

comprehensive AR protein database in a target protein database, MiCId's workflow can potentially deal with the horizontal AR gene transfer, and the presence of a few mutations in an AR protein does not prevent it from being identified provided that there are sufficient identified peptides containing no mutations. This can potentially allow the presence of few mutations occurring in the AR proteins to be identified. Overall, the target protein database used in MiCId's strategy is not too far off from the ideal target database because the proteomes of most strains under a given species share a significant number of highly homologous proteins[55,56] and the inclusion of AR proteins in a general manner takes care of the possible gain, via horizontal gene transfer, of known AR proteins.

We have used five MS/MS data sets, consisting in total of 126 HPLC−MS/MS datafiles (each containing about 20000−30000 spectra), covering 10 antibiotic-resistant bacterial strains, to evaluate the newly augmented MiCId workflow in terms of AR protein identifications. In our evaluation, AR proteins are identified at the AR protein family level, following the AR protein family classification used by the CARD database.[52,57] Identification of AR proteins is performed at the family level because of the large number of highly homologous AR proteins within most AR protein families. (Many AR proteins within the same AR family differ from each other by only one to few amino acid residues.) The high degree of protein sequence similarity makes the task of distinguishing among individual proteins beyond the AR protein family level not always possible, especially when a data-dependent acquisition mode is used in MS/MS experiments. Although identification of the exact AR protein is not always possible, obtaining identifications at the AR protein family level are enough to improve antibiotic treatments for patients suffering from bacterial infections since AR proteins within the same AR protein family are largely resistant to the same antibiotics.

In our evaluation, we have shown that MiCId's workflow has a sensitivity of approximately 85% (with an estimated lower bound of 72%) and a precision greater than 95% in the identification of AR protein families. We have demonstrated, using an MS/MS data set from samples of two human pathogens, that MiCId's workflow can be employed to investigate possible mechanisms of antibiotic resistance in bacteria. We have also shown that MiCId's workflow can provide microorganismal identification, protein identification, sample biomass estimation, and AR protein identification in 6−17 min using computer resources that are available in most desktop and laptop computers. The new MiCId version v.07.01.2021, designed to run in a Linux environment and tested under (i) CentOS Linux release 7.9.2009, (ii) Red Hat Enterprise Linux Server release 7.9, (iii) Ubuntu release 18.04.3, and (iv) Windows 10 using Oracle VirtualBox 6.1.22 running Ubuntu release 18.04.3, is freely available for download at https://www.ncbi.nlm.nih.gov/CBBresearch/Yu/downloads.html.

2. MATERIALS AND METHODS

2.1. MiCId's AR Protein Identification Algorithm.

MiCId's workflow is augmented to allow for AR protein identifications. MiCId's workflow contains procedures for taxonomic identifications, biomass estimations, and protein identifications. In the workflow, it is the protein identification part that gets augmented for the purpose of AR protein identifications. Below, we summarize MiCId's workflow and highlight the augmentations required. MiCId begins by querying a sample's MS/MS spectra in the microorganismal database, containing protein sequences from reference and representative genomes, for the identifications of microorganismal peptides; these identified microorganismal peptides are then used for taxonomic identifications via an iterative approach at each taxonomic level, and for relative taxa biomasses estimates within

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

the sample.[29,30] The proteins from the reference/representative proteomes of species identified with *E*-value ≤0.01 and *prior* ≥ 0.01 are then assembled on-the-fly for protein identification.

In the augmented MiCId, we add to the aforementioned protein database AR proteins from an AR database. Namely, in the protein identification procedure, MiCId now queries the updated protein database (combining the protein database constructed on-the-fly and the AR database) with MS/MS spectra to identify peptides for protein and AR protein identifications. (This should not be confused with the peptide identifications needed for taxonomic identifications and biomass estimates.) MiCId uses the scoring function and statistics from the database search tool RAId_DbS[49] to score peptides and for assigning statistical confidences, *E*-values, to identified peptides. Identified peptides are then used as evidence for protein identifications. See Figure 1 for an overview of MiCId's workflow.

MiCId aims to identify AR protein candidates that are globally homologous to the AR proteins already validated (e.g., proteins in an AR database). When performing protein identifications, proteins that share a large number of identified peptides are grouped as a cluster. To control the number of identified proteins, several existing methods[58] report those similar proteins as one. Adopting the same idea, we implemented this approach via two clustering procedures: (1) a peptide-centric clustering procedure and (2) a protein-similarity clustering procedure. Details regarding the clustering procedures are provided in the first section of Supplementary File S1.

2.2. MS/MS Data Sets. A total of five MS/MS data sets were used for this study. One data set, generated in-house PXD026634, is composed of 21 experimental MS/MS datafiles from samples of five bacteria strains. The other four data sets, downloaded from the ProteomeXchange Database (PD),[59] contain 105 experimental MS/MS datafiles from samples of five other bacterial strains. For seven bacterial strains used in this study, one may download their complete genomic sequences[24,60−62] and protein sequences from the National Center for Biotechnology Information (NCBI) databases.[63] In Table S1, we provide the pertinent information for each MS/MS data set.

2.2.1. In-House MS/MS Data Set. Two carbapenem-resistant *P. aeruginosa* strains were included in the study. Strain CCUG 51971 (= PA 66) was isolated from a human urine sample, at the Karolinska Hospital (Stockholm, Sweden), carrying OXA-35, OXA-488, PDC-35, and VIM-4.[60] The VIM-4 metallo-β-lactamase is responsible for the high carbapenem resistance levels (minimum inhibitory concentration (MIC) of imipenem and meropenem greater than 256 μg/mL; MIC of imipenem + ethylenediaminetetraacetic acid [EDTA] = 6 μg/mL).[60] Strain CCUG 70744 was isolated from a human sputum sample, at the Sahlgrenska University Hospital (Gothenburg, Sweden), carrying OXA-905 and PDC-8.[62,64,65]

Furthermore, one *E. coli* and two *K. pneumoniae* strains, isolated from various clinical samples at the Sahlgrenska University Hospital, carrying different β-lactamases (including extended spectrum β-lactamases, ESBL, and carbapenem resistance genes) were included in the study. *E. coli* CCUG 70745 isolated from human feces, carrying CMY-6, CTX-M-15, NDM-7,and OXA-1; *K. pneumoniae* CCUG 70742 isolated from human urine, carrying CTX-M-15, OXA-1, OXA-48, and TEM-1; and *K. pneumoniae* CCUG 70747 isolated from human wound, carrying KPC-2, SHV-200, TEM-1, and VIM-1.[62] Lyophiles of all strains were obtained from the Culture Collection of University of Gothenburg (CCUG, Gothenburg, Sweden; www.ccug.se). The strains were reconstituted on Müller-Hilton agar (Substrate Unit, Department of Clinical Microbiology, Sahlgrenska University Hospital), at 37 °C, for 24 h.

Further details regarding cultivation conditions, sample preparation, and LC−MS/MS acquisition are provided in the second section of the Supplementary File S1.

2.2.2. Downloaded MS/MS Data Sets. Four publicly available data sets, previously used in two different studies on the identification of AR proteins in bacteria, were downloaded from the ProteomeXchange Database (PXD) (http://www.proteomexchange.org/). Data set PXD004321 was taken from the study of the computational method TCUP on the identification of AR proteins.[24] This data set contains six experimental MS/MS datafiles from samples of a ESBL *E. coli* strain CCUG 62462, carrying CTX-M-15 and TEM-1;[61] the CCUG 62462 strain was grown in pure cultures without and with cefotaxime at 1000 μg/mL. Data set PXD011105, containing 35 experimental MS/MS datafiles, was taken from the study on the mechanism of antibiotic resistance of two clonal isolates (the *P. aeruginosa* strain CLJ1 antibiotic-sensitive isolate and the *P. aeruginosa* strain CLJ3 multidrug-resistant isolate obtained from the same patient at different times) grown in pure cultures with carbenicillin at 200 μg/mL.[66] Data set PXD005587, containing 24 experimental MS/MS datafiles, was taken from the investigation on proteomics changes due to antibiotic-dependent perturbations in ESBL *K. pneumoniae* strain 34618, grown in pure cultures without and with doxycycline or streptomycin.[31] Data set PXD010244, containing 40 MS/MS datafiles, was taken from the research on the mechanism of antibiotic resistance in ESBL *K. pneumoniae strain* KpV513 grown in pure cultures without and with doxycycline or streptomycin or doxycycline and streptomycin.[32]

2.3. MS/MS Data Analysis Using MiCId Workflow. All data sets were analyzed using the MiCId (v.07.01.2021) workflow.[21,29,30] The peptide-centric microorganismal database used for analysis includes all reference and representative proteomes of bacteria (12,703 strains in total) that are available in the National Center for Biotechnology Information (NCBI) database as of Feb 4, 2021. The reference proteome of one *Homo sapiens* is also included for two reasons. First, human proteins are a major component of microorganism samples when obtained directly from human hosts. Second, human proteins (mostly keratin) are also frequently identified, albeit at lower abundances, even in microorganism samples from laboratory cultures. The reference proteomes of *Bos taurus* and *Saccharomyces cerevisiae* are also included because they are present, respectively, in the Mueller Hinton Broth and in the Luria−Bertani Broth; both broth media are routinely used to grow bacterial cultures. The protein sequence Fasta files for the 12703 organisms along with the file containing taxonomic information were downloaded from the NCBI database at https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/ and at https://www.ncbi.nlm.nih.gov/Taxonomy on Feb 4, 2021. In total, 60176722 protein sequences were downloaded. As previously described,[29] to speed up MS/MS spectrum analysis, MiCId processes the protein sequences and the taxonomic file into a peptide-centric microorganismal database. The final size of the peptide-centric microorganismal database is 100 GB.

To allow for AR protein identifications, in addition to the database mentioned above, MiCId included in its search scope AR proteins from one of the following databases: ResFinder,[27] CARD,[52] or NDARO.[53,67] Users also have the option to provide

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

for MiCId their own assembled AR protein database in a Fasta file. Table S2 lists the protein identifiers for the AR proteins along with the taxonomy identifiers and scientific names of the organisms included in MiCId's databases.

While querying the database with PXD004321, PXD026634, PXD011105, PXD005587, and PXD010244 data sets, the following search parameters were employed. The digestion rules of trypsin and lys-c were assumed with up to two missed cleavage sites per peptide. The mass error tolerance of 5 ppm was set for the precursor ions and 20 ppm for the product ions except when analyzing PXD011105, PXD005587, and PXD010244 (10 ppm for the precursor ions). For PXD004321 and PXD026634, cysteines were unmodified, and for PXD011105, PXD005587, and PXD010244 iodoacetamide was used as the reduction agent, changing the molecular mass of every cysteine from 103.00919 to 160.030647 Da. RAId's Rscore, using b- and y-ions as evidence, was used to score peptides. The statistical significance of each peptide was assigned via RAId_DbS's theoretically derived peptide score distribution.[49] The largest (cutoff) *E*-value for a peptide to be reported was set to 1. For taxa identifications at the genus level and lower, all microorganisms under the genus *Shigella* were excluded from consideration to avoid classification ambiguity because some researchers have argued that taxonomically *Shigella* should be classified under *Escherichia coli*.[68]

3. RESULTS AND DISCUSSION

In this section, we present the evaluation of MiCId's workflow in identifying AR proteins. First, we use 126 experimental MS/MS datafiles to assess MiCId's AR protein identification strategy. Second, we estimate the sensitivity of AR protein identifications via MiCId using 27 experimental MS/MS datafiles from samples of six antibiotic resistant bacteria strains (from three species included in the pathogen priority list of the World Health Organization[69] for antibiotics research and development), cultured with and without antibiotics. Third, using 35 experimental MS/MS datafiles from samples of two human pathogens, we employ MiCId's workflow to investigate possible mechanisms of antibiotic resistance.

Following the AR protein classification used by the CARD database,[52,57] one finds that there are large numbers of highly homologous AR proteins within most AR protein families and expects this family level crowdedness to remain as AR protein databases continue to grow. As an illustration, we note that each AR protein within the β-lactamase family has very homologous sequences within the family: if one takes an AR protein as the query to align with each of the rest of the AR proteins in the β-lactamase family, for the best pairwise alignment, there are many high score alignments and the best of which has an average length normalized BLAST bit-score $\geq$2. This is shown in Table S3. The length normalized BLAST bit-score is defined as the BLAST bit score divided by the length of the longer of the two sequences aligned. As illustrated in Figures S1 and S2, a good cutoff for length normalized BLAST bit-score is 1.6.

The high degree of similarity for AR proteins in the same family makes distinguishing among AR proteins at finer-than-family level not always possible, particularly when data acquisition in the MS/MS experiment is untargeted. For these reasons, during our evaluation, identified AR proteins are counted as true positives and false positives at the AR protein family level. For example, assume a bacteria strain contains OXA-1 and OXA-48 proteins, leading to two proteins in the OXA family; if during the analysis MiCId identifies three OXA proteins, then the two best ranking OXA proteins identified are counted as true positives and the remaining OXA protein is counted as a false positive. For some AR protein families that are not yet overly represented/crowded in the database, correct identification can be achieved at finer-than-family when closely related homologous proteins are present in the database. For example, for AR proteins from aminoglycoside families, families that are not overly represented in the database, we observed correct identifications for these AR proteins not only at the family level but also at the isoenzyme level,[70,71] which is a finer level than the family level. Of course, if the target family (to which the query protein belongs) is too much underrepresented in the database, either no identification is made or misidentifications of the AR protein families occur.

3.1. Evaluation of MiCId's AR Protein Identification Strategy. MiCId's strategy for AR protein identifications is to first identify species in a microorganismal database and then identify AR proteins in a target protein database composed of proteins from the reference/representative proteomes of confidently identified species and AR proteins from a high-quality AR database.[27,52,53] MiCId's strategy capitalizes on microorganismal identifications at the species level because high confidence microorganism identifications at taxonomic levels lower than species become challenging because of the lack of discriminative peptides among the ones identified[30,72] when using the routinely employed high-resolution data-dependent acquisition mode in MS/MS experiments.[73] In principle, more advanced MS/MS experiments such as targeted MS/MS using selected reaction monitoring (SRM) or parallel reaction monitoring (PRM) can be used for taxonomic identification below the species level by targeting peptides that are unique to taxa at lower taxonomic levels.[72,74−77] However, a limitation of such approaches is that they can only be employed for the identifications of a microorganism within a small, predetermined set of microorganisms. Another reason for employing MiCId's strategy has to do with the trustworthiness in annotation of the taxonomic database for taxonomic levels below species.[78−81] It is important to mention that although for this study we only included the proteins from strains that are labeled as reference and representative in the microorganismal database, as these are proteins from higher quality genomes,[54,82,83] MiCId's workflow is not limited to microorganismal databases composed of only reference and representative genomes, and it can perform microorganismal identifications beyond the species level.

When, for the purpose of protein identifications, selecting a proteome as the representative for a confidently identified species, MiCId relies on a heuristic because under a given species there could be many strains and priority for each has to be established. The heuristic gives strains that are labeled as *reference* first priority and *representative* second priority. Information about reference and representative strains is taken from the RefSeq and GenBank assembly summary files downloaded from the NCBI. If there is more than one reference strain or representative strain for a given species, the strain with the larger number of proteins is selected. When a species has neither reference strain nor representative strain, the proteome from the strain, under that species, with the larger number of proteins is selected. The rule of assigning high priority to the proteomes from reference strains and representative strains is applied because these are proteome assemblies of higher quality and importance that have been curated by the NCBI staff and are to be used as anchors for the analysis of closely related proteomes within the same taxonomic group.[54]

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

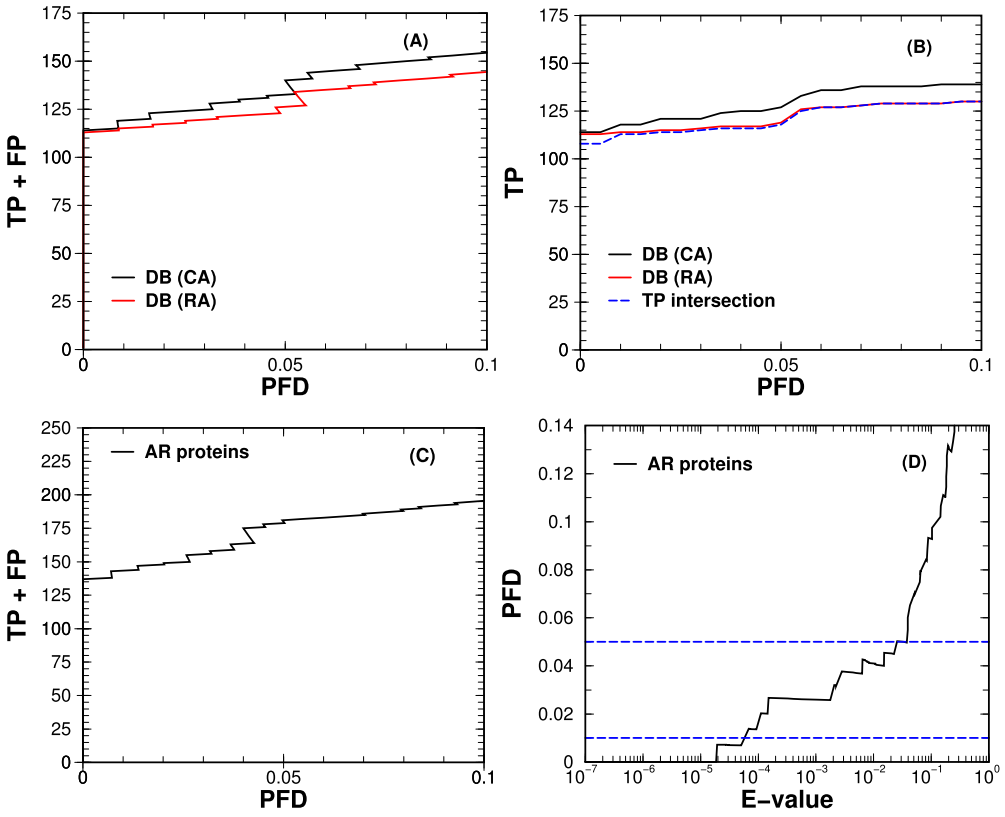


Figure 2. MiCId workflow evaluation. Let us mention here again that the abbreviations CA and RA refer, respectively, to target databases composed of proteins from the correct strain plus the chosen AR database and from reference strains plus the chosen AR database. Panels A and B display PFD curves when querying the CA and RA with 62 experimental MS/MS datafiles. Panel A shows that the PFD curves from searching in CA and RA are comparable. Panel B shows that there are 131 true positive antibiotic resistance (AR) proteins identified in common in CA and RA. The PFD values in panels C and D were obtained from querying RA with 126 experimental MS/MS datafiles. Panel C also indicates that 180 AR proteins are identified at the 5% PFD level. Panel D shows that using an *E*-value cutoff of 0.01 the identification of AR proteins can be controlled at the PFD level smaller than 5%. The abbreviations TP, FP, and PFD refer, respectively, to true positive, false positive, and proportion of false discoveries.

For identifications of AR proteins, the ideal target protein database would include all of the protein sequences obtained directly from the strains present in the biological sample and with AR proteins unambiguously annotated. From an MS/MS-based proteomics viewpoint, such a database is unattainable even if strain level identification is achieved. MS/MS-based proteomics approaches rely on databases such as the ones at the NCBI to obtain protein sequences for yet-to-be-identified strains. A target protein database constructed using this procedure would still be an approximation to the ideal target protein database because the strains present in the biological sample could have acquired new proteins via horizontal gene transfer and mutations through rapid multiplication and environmental pressure.[50,51] By including a comprehensive AR protein database in the target database, MiCId can potentially deal with the horizontal AR gene transfer; with the clustering procedure, a few mutations of an AR protein do not prevent it from being identified provided that there are sufficient identified peptides containing no mutations. However, lacking a complete AR protein database encoding[84] all existent mutations, MiCId cannot pinpoint the mutation sites and their amino acid polymorphisms. MiCId can potentially allow the presence of a few mutations occurring in the AR proteins to be identified. Overall, the target protein database used in MiCId's strategy is not a bad approximation to the ideal target database because the proteomes for most strains under a given species shared a significant number of homologous proteins[55,56] and include AR proteins in a global manner, covering the possible acquisition, via horizontal gene transfer, of known AR proteins.

To evaluate MiCId's strategy for identifying AR proteins, we prepared two sample-specific target protein databases and queried them with the same MS/MS datafiles from specific samples. The first target protein database is composed of proteins from the reference proteome of the species present in the sample and AR proteins from the ResFinder database, referred to here as reference strains plus AR database (RA). The other target protein database is composed of proteins from the proteome of the true strain present in the sample and AR proteins from the ResFinder database, referred to here as correct strain plus AR database (CA). Table S4 contains the protein identifiers for each protein used to generate both versions of the sample-specific target protein databases. Plotted in panels A and B of Figure 2 are the PFD curves from querying the RA and CA databases with 62 experimental MS/MS datafiles from samples of seven strains. There are in total 14 target protein databases, two for each of the seven strains that have complete genome sequence available, used in generating panels A and B of Figure 2. The PFD curves in panel A of Figure 2 show that using the target protein databases RA and CA produced comparable PFD curves. Furthermore, the curves in panel B of Figure 2 show that using RA and CA databases yields 131 common AR protein identifications. What was not shown is that there are 12 AR protein identifications, covering five AR protein families, not shared: there are eight PDC protein family identifications and

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

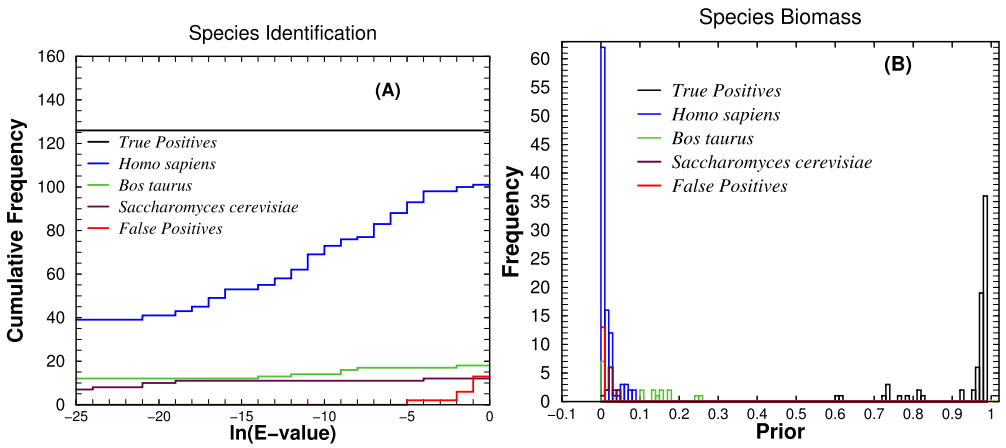


Figure 3. Species level composition for samples 1−126. Plotted in panel A are the cumulative frequency of the number of true positives species, *H. sapiens*, *B. taurus*, *S. cerevisiae*, and false positives species versus the natural logarithm of the *E*-value of all species identified with *E*-value ≤ 1. Plotted in panel B are the histograms of priors for the true positive species, *H. sapiens*, *B. taurus*, *S. cerevisiae*, and false positive species. Among the true positives: *E. coli* is identified 10 times, *K. pneumoniae* 72 times, and *P. aeruginosa* 44 times. Among the 13 false positives: *Algisphaera agarilytica* is identified 1 time, *Cerasicoccus arenae* 4 times, *Chlorobium tepidum* 2 times, *Desulfosporosinus orientis* 1 time, *Fervidobacterium thailandense* 1 time, *Ktedonosporobacter rubrisoli* 1 time, *Streptococcus thermophilus* 1 time, and *Thiofilum flexile* 2 times. *H. sapiens* is identified 103 times, *B. taurus* 18 times, and *S. cerevisiae* 12 times. To control the proportion of false discoveries below 5% only species identified with an *E*-value ≤ 0.01 (ln(*E*-value) = −4.6) and *prior* ≥ 0.01 are considered true positives with high confidence. When employing the recommended cutoffs of *E*-value ≤ 0.01 and *prior* ≥ 0.01 MiCId still identifies all true positives with no false positives.

the one ant(3") family identification present in the list identified using the CA database but absent from that using the RA database; on the other hand, only one TEM family identification, one OXA family identification, and one ARR family identification are found using the RA database. Multiple PDC family identifications are found using both databases: 23 identifications using the CA database and 15 identifications using the RA database. The identification rate of PDC protein family in the CA database is higher because it contains the correct PDC protein PTC38756.1, which belongs to the CLJ1 strain, even though this protein is not yet included in the ResFinder database. In addition, in the ResFinder database the PDC family—containing only four PDC proteins: AAM08942.1, ACQ82815.1, ACQ82807.1, and AAM08945.1—is under-represented, making it difficult to identify the correct PDC using the ResFinder database since even the most homologous PDC protein (AAM08942.1) and the correct PDC protein (PTC38756.1) differ by more than 50 amino acid residues. The discrepancy in true positives in the other four AR protein families is also mainly caused by composition difference of the two target databases. Table S5 contains pertinent information on all the identified proteins/families in both databases.

Panel C of Figure 2 shows that there are 180 AR protein family identifications at the 5% PFD level when all 126 MS/MS data files are analyzed. Panel D of Figure 2 shows that when an *E*-value cutoff of 0.01 is used the identification of AR proteins can be controlled at the 5% PFD level. On the basis of this result, in order to control the false positives at around the 5% PFD level, only AR protein family identifications with *E*-values below 0.01 are deemed true positives with high confidence by MiCId. Table S6 has the list of all identified proteins for all 126 MS/MS data files. Because the 126 datafiles used have been annotated with true positives, we are able to display the "theoretical" PFD values in Figure 2 to show the retrieval effectiveness. However, in real experimental data analyses, the PFD has to be estimated as the true positives and false positives are not known beforehand. In the third section of Supplementary File S1, we show how to estimate the PFD values via *E*-values. The closeness between the "theoretical" PFD and estimated PFD is also shown in Figure S3. Even though the PFD can be estimated, control of PFD does not prioritize the proteins that meet the PFD cutoff. For this reason, we find that *E*-values, when assigned accurately, provide more useful information. Not only it can be used to infer the expected number of false positives, hence, type-I error control, it can also be used to prioritize the proteins meet a PFD cutoff. We also like to stress that MiCId searches the database in a single pass with the PFD computed via the accurate *E*-values reported;[40,49,85] it does not use multipass target-decoy heuristics. The latter was designed with the intent to amplify the identification rates but, unfortunately, violates the statistical foundations, of the target-decoy approach.[86−89]

As mentioned above, a requirement for MiCId's strategy to work is that it must have accurate species-level identification. MiCId achieves accurate microorganism identification with trustworthy confidence assignments by properly computing for every identified taxon an *E*-value and a *prior* probability.[29,30] For a quality score *S*, the *E*-value reflects the expected number of random taxa with scores the same as or better than *S*.[40] A taxon's *prior* probability is the probability for an identified taxon to emit any evidence peptide which can also be viewed as that taxon's protein biomass up to an overall proportionality constant as described earlier.[30] Therefore, identified taxa with small *E*-values and large priors are more likely to be present in the sample. As we have demonstrated, MiCId can control the PFD below 5% by calling true positives only identified taxa with *E*-values ≤ 0.01 and *prior* ≥ 0.01.[29,30] In addition, MiCId employs an iterative approach for taxa identification at each taxonomic level; only taxa identified at the upper taxonomic level are considered for the next level identifications.[29,30] As shown in Figure 3, when considering all identifications with *E*-values ≤ 1, MiCId identifies for each of the 126 samples the correct species with only 13 false positives overall. Interestingly, MiCId also identifies *H. sapiens* in 103 samples, *B. taurus* in 18 samples, and *S. cerevisiae* in 12 samples. *H. sapiens* are identified using evidence peptides from keratin proteins detected in the samples.

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

Table 1. Identification Results of β-Lactamase Proteins from Culture Samples of Six Antibiotic-Resistant Strains Cultivated with and without β-Lactam Antibiotics[a]

E. coli CCUG 62462						
AR Protein	SN-1 NA	SN-2 NA	SN-3 NA	SN-4 CTX 1mg/ml	SN-5 CTX 1mg/ml	SN-6 CTX 1mg/ml
CTX-15	✓	✓	✓	✓	✓	✓
TEM-1	✓	✓	✓	✓	✓	✓
E. coli CCUG 70745						
AR Protein	SN-7 ETP 56μg/ml	SN-8 ETP 56μg/ml	SN-9 NA	SN-10 NA		
CTX-15	✓	✓	✓	✓		
CMY-6	✓	✓	✓	✓		
NDM-7	✓	✓	✓	✓		
OXA -1	✓	✓	✓	✓		
P. aeruginosa CCUG 51971						
AR Protein	SN-11 NA	SN-12 MEM 8μg/ml	SN-13 MEM 32μg/ml	SN-14 MEM 128μg/ml	SN-15 MEM 256μg/ml	
OXA-35	✓	✓	✓	✓	✓	
OXA-488	✓	✗	✗	✗	✗	
VIM-4	✓	✓	✓	✓	✓	
PDC-35	✗	✗	✗	✓	✓	
P. aeruginosa CCUG 70744						
AR Protein	SN-16 NA	SN-17 MEM 2μg/ml	SN-18 MEM 4μg/ml	SN-19 MEM 8μg/ml		
PDC-8	✓	✓	✓	✓		
OXA-905						
K. pneumoniae CCUG 70742						
AR Protein	SN-20 ETP 21μg/ml	SN-21 ETP 21μg/ml	SN-22 ETP NA	SN-23 ETP NA		
CTX-15	✓	✓	✓	✓		
OXA-1						
OXA-48	✓	✓	✓	✗		
TEM-1	✓	✓	✓	✓		
K. pneumoniae CCUG 70747						
AR Protein	SN-24 ETP 28μg/ml	SN-25 ETP 28μg/ml	SN-26 ETP NA	SN-27 ETP NA		
KPC-2	✓	✓	✓	✓		
VIM-1	✓	✓	✓	✓		
SHV-200	✓	✓	✗	✓		
TEM-1	✓	✓	✓	✗		

[a]Cells in green color and marked with a checkmark indicates that the protein was identified with an *E*-value $\leq$ 0.01, indicating high confidence in the identification. Cells in yellow and marked with a checkmark indicate that the protein was identified with an 0.01< *E*-value $\leq$ 1, indicating low confidence in the identification. Cells with X indicate that the protein was not identified for that sample number (SN); cells with no mark indicate that the protein was not identified in that data set; CTX, cefotaxime; ETP, ertapenem; MEM, meropenem; NA, no antibiotic. Cases of no identification have no mark.

Keratin proteins are a common contaminant to mass spectrometry experiments, usually originating from skin and hair as well as dust, clothing, and latex gloves. The identification of *B. taurus* and *S. cerevisiae* in some of the samples is expected as they are present in the broth medium used to grow the bacterial cultures.

When imposing the recommended cutoffs, *E*-values $\leq$ 0.01 and *prior* $\geq$ 0.01, to control the PFD below 5%, MiCId still identifies correctly the true positive species out of each of the 126 samples. This is because, as shown in Figure 3, all of the true positives are identified with a much lower *E*-value than 0.01 and much larger *prior* than 0.01. However, with the recommended cutoffs, *H. sapiens* is identified now in 41 samples, *B. taurus* in 11 samples, *S. cerevisiae* in 11 samples, and no false positives. In terms of the *prior*, reflecting the taxon's relative protein biomass, one would expect it to be very close to 1 for true positive species identified, given that the samples are each assumed to contain a single microorganism. The main reason that it deviates from 1 is because out of the 126 samples, when one only imposes *E*-values $\leq$ 1 for reporting identification, 105 samples have, in addition to the underlying microbe, identifications matching some of the following three organisms: *H. sapiens*, *B. taurus*, and *S. cerevisiae*. For these 105 samples, as shown in panel B of Figure 3, *H. sapiens* contributes to the overall protein biomasses with *prior* values ranging from 0.00076 to 0.085 with an average value of 0.016; *B. taurus* has *prior* values ranging from 0.0019 to 0.25 with an average value of 0.1; *S. cerevisiae* has *prior* values ranging from 0.0011 to 0.048 with an average value of 0.026. These non-zero *prior* values for *H. sapiens*, *B. taurus*, and *S. cerevisiae* cause the observed deviation of the *prior* value from 1 for the TP. Table S7 contains pertinent information on the identified species for each sample.

It is important to mention that the taxa identification results reported by MiCId are not filtered by using the recommended cutoff to avoid incidental false negatives. MiCId reports the complete list of identified taxa using a color-coded scheme. Identified taxa passing the recommended cutoffs, *E*-values $\leq$0.01 and *prior* $\geq$ 0.01, are highlighted in green for high-confidence in being a true positive; taxa identified with an *E*-value $\leq$ 1 and *prior* $\geq$ 0.001 are highlighted in yellow for low confidence in being a true positive, and taxa identified with an *E*-value > 1 or

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

prior < 0.001 are highlighted in red for no-confidence in being a true positive.

3.2. Estimate for Sensitivity of AR Protein Identifications via MiCId's Workflow. Having computational methods that can correctly identify bacteria and also their AR proteins is among the most important research fronts for fighting infections. We demonstrate the usefulness of MiCId's workflow in serving as such a computational method in this subsection and next. We use datafiles from some bacteria containing β-lactamase proteins as examples for the reasons listed below. First, β-lactam antibiotics are the most prescribed class of antibiotic to fight infections globally;[90] second, in the United States, about 65% of the antibiotics prescribed are β-lactam antibiotics.[91] Of special importance in this class of antibiotic are carbapenems. Carbapenems have a broad spectrum of activity and are usually used as the last-line of the defense for seriously ill patients suspected of harboring resistant bacteria.[92] Evidently, correct identifications of carbapenem resistance can help significantly in fighting infections. In addition, β-lactamase proteins can be harbored by plasmid, and when this occurs they can be easily transmitted into different bacteria cells, introducing resistance to the bacteria.[90,93,94]

To estimate the sensitivity of MiCId's workflow on the identification of AR proteins, we used 27 MS/MS experimental datafiles from six antibiotic-resistant bacterial strains (from three species included in the pathogen priority list of the World Health Organization[69] for antibiotics research and development), cultured with and without β-lactam antibiotics. The three β-lactam antibiotics used belong to two classes of antibiotics: belonging to the cephalosporin class is cefotaxime and belonging to the carbapenem class are ertapenem and meropenem. Each of the bacterial strains carries between two and four predicted β-lactamase proteins and shows resistance to a variety of antibiotics.[60−62,65,95] β-Lactamase proteins for these strains were computationally predicted using ResFinder.[27] Table S8 provides a protein-centric view. This table lists for each predicted β-lactamase the strains containing it and the names of the β-lactam drug classes it resists. For the purpose of estimating the sensitivity value, we view each possible β-lactamase identification per experiment as a different event. Summing the numbers of possibly identifiable β-lactamase proteins from each of the 27 experiments, one obtains a total of 88 potential true positives. This may be viewed as the maximum set of the true positives. An avid reader may ask what happens if some AR proteins, in this case β-lactamase proteins, are missed from the database. When that happens, because these proteins will never be identified, they do not contribute counts to either the numerator or the denominator while the sensitivity value is computed. Hence, for the purpose of assessing the sensitivity, one does not need to worry about AR proteins that are not included in the database. On the other hand, a predicted AR protein may never be observed because it is usually expressed in low abundance or it is not even a true protein in the corresponding microorganism's proteome. When this is the case, it becomes inappropriate to use the maximum TP set as the TP set for the purpose of estimating the sensitivity value.

When using all 88 possible identifications as the TP set, one obtains a sensitivity value of 72.7% (64/88). This may be viewed as the lower bound of the sensitivity of MiCId's workflow. If one excludes from the TP set the β-lactamases—OXA-1 in *K. pneumoniae* CCUG 70742, OXA-488 in *P. aeruginosa* CCUG 51971, and OXA-905 in *P. aeruginosa* CCUG 70744—that were never confidently observed in any of the corresponding experiments considered, one obtains a sensitivity value of 85.3% (64/75). This sensitivity value may be viewed as the typical sensitivity value while employing MiCId's workflow.

Table 1 shows the identification results of β-lactamase protein families for all the 27 MS/MS experiments. Displayed in Table 1 are 64 identifications with *E*-value $\leq$ 0.01 highlighted in green and marked with a checkmark, six identifications with 0.01 < *E*-value $\leq$ 1 highlighted in yellow and marked with a checkmark, five cases of missed identification (while identified in other samples) marked with an X, and 12 cases of no identification with no marks.

For bacterial cultures exposed to an antibiotic, one would expect the bacteria to express some of its AR proteins at high levels.[96,97] MiCId's workflow does identify, except for OXA-1, OXA-488, and OXA-905, all of the predicted β-lactamase proteins. The AR protein OXA-1 is copresent with OXA-48, CTX-15, and TEM-1 in the genome of *K. pneumoniae* CCUG 70742; OXA-488 is present along with OXA-35, VIM-4, and PDC-35 in the genome of *P. aeruginosa* CCUG 51971; and OXA-905 is copresent with PDC-8 in the genome of *P. aeruginosa* CCUG 70744. For *K. pneumoniae* CCUG 70742, MiCId's workflow identified OXA-48 in three samples via OXA-232, OXA-199, and OXA-548 as these three proteins are highly homologous to OXA-48 and have length-normalized BLAST bit-scores of 2.04, 2.05, and 1.69, respectively. For *P. aeruginosa* CCUG 51971, MiCId's workflow identified OXA-35 in five samples with high confidence via OXA-19, OXA-101, OXA-35, OXA-147, and OXA-240 as these five proteins are highly homologous to OXA-35 and have length-normalized BLAST bit-scores of 2.04, 2.03, 2.05, 2.03, and 1.98, respectively. The complete list of identified AR proteins can be found in Table S6.

MiCId's identification results for *P. aeruginosa* CCUG 51971 and *P. aeruginosa* CCUG 70744 correlate well with a previous study showing that, in the model strain *P. aeruginosa*—PAO1 the gene of the OXA-50-like oxacillinase—is expressed at relatively low levels and is not inducible by β-lactams, while the gene of bla_{PDC}, also expressed at relatively low levels usually, is strongly induced by β-lactams.[98] This could be the reason why MiCId did not detect OXA-488 in *P. aeruginosa* CCUG 51971 and OXA-905 in *P. aeruginosa* CCUG 70744 but detected PDC-35 in *P. aeruginosa* CCUG 51971, albeit only at the highest concentrations of meropenem. There are also several experimental reasons ranging from digestion enzyme, data-dependent acquisition mode selection, protein expression level, as well as nonoptimal liquid-chromatography separation that can be used to explain why some of the β-lactamase proteins were not identified or were not confidently identified. To further validate that the missed identification of β-lactamase proteins was not due to MICId's inability, we analyzed all 27 MS/MS experimental datafiles using the Proteome Discoverer software (version 2.4), and the results obtained, displayed in Table S9, are in agreement with the MiCId results.

This assessment shows that MiCId's workflow has a typical sensitivity value around 85% (and with a lower bound at about 72.7%), suggesting that it is a useful tool for the detection of AR proteins.

3.3. Using the MiCId Workflow to Investigate Possible Mechanisms of Antibiotic Resistance. We demonstrate here how the MiCId workflow may aid in the investigation of the possible mechanisms of antibiotic resistance of a human pathogen, *Pseudomonas aeruginosa* strain CLJ3, and compare the mechanism suggested by using MiCId with published results.[66] *P. aeruginosa* strains were obtained from a patient

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

having hemorrhagic pneumonia but treated unsuccessfully with antibiotics. Strain CLJ1, sensitive to antibiotics, was isolated before antibiotic therapy started; 12 days after antibiotic therapy started, as the patient conditions worsened, strain CLJ3 was isolated. A multiomics approach was used to understand the process of antibiotic resistance development in CLJ3. Genomics data shows that the genome of CLJ3, when compared to the genome of CLJ1, has acquired several genetic modifications that could have contributed to phenotypic changes. Genomics data shows that antibiotic resistance of CLJ3 is probably linked to interruption-causing insertions detected in genes *oprD* and *ampD*.[66]

For each strain, proteomics samples comprising proteins contained in the whole-cell (W), inner and outer membranes (M), and secretome (S) were collected and used for MS/MS analysis.[66] The CARD database was used as the input AR protein database in MiCId's workflow as it contains proteins belonging to multidrug efflux pumps.[99] The suggested mechanisms of antibiotic resistance for CLJ3 by using MiCId's workflow agrees with the published results.[66] Comparing the AR proteins identified in membrane samples from CLJ3 and CLJ1, one notes that CLJ3 does not express the outer membrane protein *oprD* and is overexpressing the β-lactamase PDC. Lack of the outer membrane protein *oprD*, caused by interruption of *oprD* gene, makes the cell impermeable to most antibiotics in the β-lactam class.[66,100,101] Interruption in the *ampD* gene brings about the overexpression of bla_{PDC} as the *ampD* gene is responsible for the regulation of bla_{PDC}.[102,103]

Table S10 contains the identifiers of AR protein families for each strain. Table S10 also shows that in agreement with the previous study on the mechanism of antibiotic resistance for CLJ3 was only obtained for the membrane samples. One obvious reason is that one expects to find higher concentration of *oprD* proteins in the membrane extract and of β-lactamase proteins in the periplasm, which is the cellular component between the inner- and outer-membrane of Gram-negative bacteria, and thus can often be a component contaminant for the membrane samples.[104,105] This accentuates the necessity of sample extraction selection and sample fractionation when investigating possible mechanisms of antibiotic resistance.[105−107]

We further used principal component analysis (PCA) to demonstrate the reproducibility of MiCId's workflow. The vector component for each sample was set to be $\ln(1/E\text{-value})$ of the identified AR protein family. AR proteins not identified in a given sample was assigned the *E*-value 100, yielding a vector component value of −4.605. For each sample, the components are further scaled to have norm 1. Figure 4 shows tight clusters for samples derived from whole-cell and membrane for each strain; for the secretome samples, data points are not as close, indicating that the secretome might not be suitable for studying AR proteins. The results from principal component analysis validate the reproducibility of MiCId's workflow in AR protein identifications as shown by the tight sample clusters in Figure 4.

3.4. Execution Time of MiCId's Workflow. With speed a main consideration, MiCId was written in C++ and its routines for organism and protein identifications were implemented using parallel programming. Hence, MiCId allows users to specify the desired number of cores for each job. Using 28150 MS/MS spectra to query two databases of sizes 100 GB (12703 organisms) and 20 GB (3868 organisms), we measured the execution time of MiCId's workflow in performing organism identification, biomass estimation, and protein identifications.

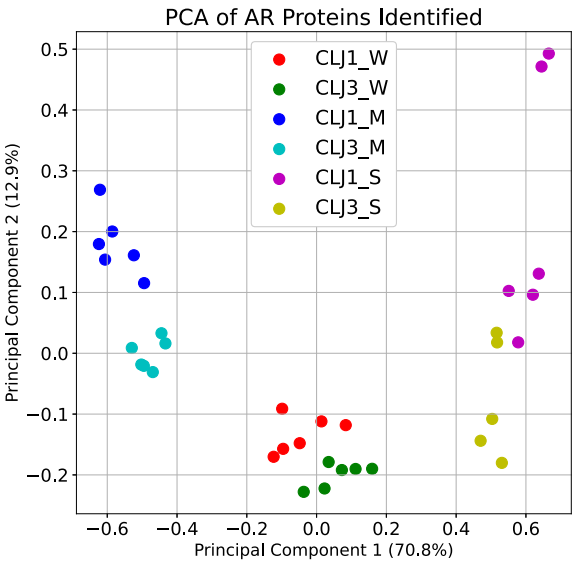


Figure 4. Principal component analysis (PCA) for antibiotic resistance (AR) protein families identified by MiCId's workflow. Included in the PCA are 35 identification results, each from an experiment whose sample contains either *P. aeruginosa* strain CLJ1 or *P. aeruginosa* strain CLJ3 with proteins collected from whole-cell (W), membrane (M), or secretome (S). Also revealed in the plot, there are only five experimental replicates for the combination CLJ3-S, the other five combinations each has six experimental replicates. This brings the total number of experiments included to 35.

Figure 5 shows that in the 20 GB database takes about 13 min with four cores and reduces to around 6 min with 16 cores. On the other hand, the execution time in the 100GB database ranges from 17 min (with four cores) down to 7 min (with 16 cores). Our results indicate that when the database size increases by a factor of 5.0, the execution time increases only by a factor of about 1.2 (using 16 cores). This reflects the scalability of MiCId in handling large databases. Figure 5 also shows that the

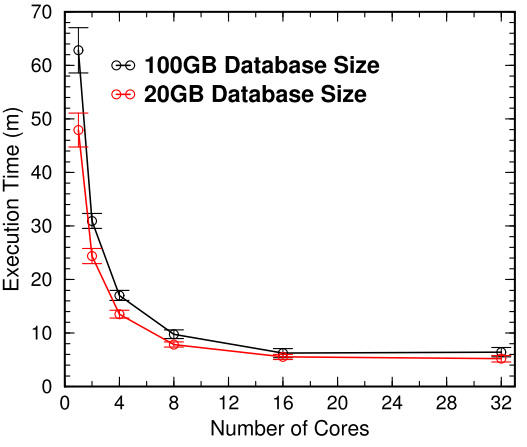


Figure 5. Average execution time, in minutes, of MiCId's workflow in performing organism identification, biomass estimation, and protein identifications in a 100GB (containing 12703 organisms) database and a 20GB (containing 3868 organisms) database. There are 28150 MS/MS spectra used as the queries. Results from using various number of cores are displayed. MiCId's workflow execution time performance was carried-out in a computer running the operating system CentOS Linux release 7.9.2009 and containing 32 Intel(R) Xeon(R) central processing units (CPUs) with a clock speed of 2.60 GHz.

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

execution time reduction reaches a plateau at around 16 cores. This is because the C++ routine used to compute statistical significance for identified organisms and proteins is not yet parallelized, incurring a constant time cost. Table S2 contains the taxonomic identifiers for all the organisms in the 100 and the 20 GB databases as well as the identifiers for proteins taken from the ResFinder database.

MiCId's workflow execution time performance was measured using a computer running the operating system CentOS Linux release 7.9.2009 and containing 32 Intel(R) Xeon(R) central processing units (CPUs) with a clock speed of 2.60 GHz. More information about the operating system and CPUs used is provided in Table S11.

4. CONCLUSION

Fast and accurate identification of pathogenic bacteria along with the identification of AR proteins is of paramount importance for patient treatments and public health. The newly augmented MiCId workflow was designed to achieve this important goal by identifying AR proteins when processing MS/MS data acquired in high-resolution mass spectrometers. The augmented workflow of MiCId also fills the need for having mass spectrometry-based workflow for identifying bacteria along with AR proteins. We have shown in section 3.1 that the strategy employed by the MiCId workflow for identifying AR protein yields sensible results. The MiCId workflow identifies 93.5% (131/140) of the AR proteins that are also identified if the target protein database used is composed of protein sequences from the correct strain. Results from our AR protein identification assessment show that MiCId's workflow has a sensitivity of 85% (with a lower bound at about 72.7%) and a precision of 95% when the *E*-value cutoff of 0.01 is used to control the number of false positives. Being fast, yielding sensible results, and having high sensitivity and high precision, MiCId is shown to be a valuable tool for identification of bacteria and their AR proteins. However, limitations to the current MiCId workflow remain. Even though the relative biomasses among multiple microbes present in a sample can be provided, MiCId does not yet provide quantification of individual proteins; although MiCId's AR protein identification allows few mutations, the impossibility of having a *complete* AR protein database prevents MiCId from pinpointing the mutation sites and types. Nevertheless, while the latter limitation cannot be circumvented with proteomics workflow alone, we do plan to address the former limitation in the near future.

The augmented workflow of MiCId is a self-contained tool capable of performing microorganism identification, protein identification, biomass estimation, and AR protein identification in minutes using limited amount of computer resources available in most desktop and laptop computers. MiCid's workflow was tested under (i) CentOS Linux release 7.9.2009, (ii) Red Hat Enterprise Linux Server release 7.9, (iii) Ubuntu release 18.04.3, and (iv) Windows 10 using Oracle VirtualBox 6.1.22 running Ubuntu release 18.04.3. Having a user-friendly graphical user interface, the new MiCId version (v.07.01.2021) for Linux environment is freely available for download at https://www.ncbi.nlm.nih.gov/CBBresearch/Yu/downloads.html.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jasms.1c00347.

Detailed information about the protein-clustering procedure used in MiCId (section 1); experimental steps used to generate the in-house MS/MS data sets (section 2); procedure to estimate the expected PFD (PFD_E) from *E*-values (section 3) and the agreement between the "theoretical/ideal" PFD and PFD_E (Figures S3) (PDF)

BLAST bit-score and length-normalized BLAST bit-score histograms (Figure S1); length-normalized BLAST bit-score cutoff learning (Figure S2); information about MS/MS files (Table S1); list of organisms and proteins used to build MiCId's microorganismal databases and protein databases (Table S2); average similarity between β-lactamase protein families (Table S3); list of protein sequence identifiers for the correct bacteria strains and for the reference/representative strains (Table S4); list of antibiotic resistance proteins identified by MiCId's workflow for sample numbers 1−62 (Table S5); list of antibiotic resistance proteins identified by MiCId's workflow for sample nos. 1−126 (Table S6); species-level identification for sample nos. 1−126 (Table S7); β-lactamase proteins and their target β-lactam drug classes for the six strains cultivated with β-lactam used in our study (Table S8); list of antibiotic resistance proteins identified by Proteome Discoverer software version 2.4 (Thermo Fisher Scientific) for sample nos. 1−27 (Table S9); list of antibiotic resistance proteins identified by MiCId;s workflow for sample nos. 28−62 when using as database proteins from the reference/representative strain plus proteins from the CARD database (Table S10); information about the computer operating system and CPUs used to measure MiCId's workflow execution time (Table S11) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Yi-Kuo Yu − *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, United States;* orcid.org/0000-0002-6213-7665; Email: yyu@ncbi.nlm.nih.gov

Authors

Gelio Alves − *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, United States*

Aleksey Ogurtsov − *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, United States*

Roger Karlsson − *Department of Infectious Diseases, Sahlgrenska Academy, University of Gothenburg, 40530 Gothenburg, Sweden; Department of Clinical Microbiology, Sahlgrenska University Hospital, 40234 Gothenburg, Sweden; Center for Antibiotic Resistance Research (CARe), University of Gothenburg, 40016 Gothenburg, Sweden; Nanoxis Consulting AB, 40234 Gothenburg, Sweden*

Daniel Jaén-Luchoro − *Department of Infectious Diseases, Sahlgrenska Academy, University of Gothenburg, 40530 Gothenburg, Sweden; Center for Antibiotic Resistance Research (CARe), University of Gothenburg, 40016 Gothenburg, Sweden; Culture Collection University of Gothenburg (CCUG), Sahlgrenska Academy of the University of Gothenburg, 40234 Gothenburg, Sweden*

Beatriz Piñeiro-Iglesias − *Department of Clinical Microbiology, Sahlgrenska University Hospital, 40234*

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931

Gothenburg, Sweden; Center for Antibiotic Resistance Research (CARe), University of Gothenburg, 40016 Gothenburg, Sweden

Francisco Salvà-Serra − *Department of Infectious Diseases, Sahlgrenska Academy, University of Gothenburg, 40530 Gothenburg, Sweden; Department of Clinical Microbiology, Sahlgrenska University Hospital, 40234 Gothenburg, Sweden; Center for Antibiotic Resistance Research (CARe), University of Gothenburg, 40016 Gothenburg, Sweden; Culture Collection University of Gothenburg (CCUG), Sahlgrenska Academy of the University of Gothenburg, 40234 Gothenburg, Sweden; Microbiology, Department of Biology, University of the Balearic Islands, 07122 Palma de Mallorca, Spain*

Björn Andersson − *Bioinformatics Core Facility at Sahlgrenska Academy, University of Gothenburg, 40530 Gothenburg, Sweden*

Edward R. B. Moore − *Department of Infectious Diseases, Sahlgrenska Academy, University of Gothenburg, 40530 Gothenburg, Sweden; Department of Clinical Microbiology, Sahlgrenska University Hospital, 40234 Gothenburg, Sweden; Center for Antibiotic Resistance Research (CARe), University of Gothenburg, 40016 Gothenburg, Sweden; Culture Collection University of Gothenburg (CCUG), Sahlgrenska Academy of the University of Gothenburg, 40234 Gothenburg, Sweden*

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.1c00347

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the administrative group of the National Institutes of Health Biowulf Cluster, where all the computational tasks were carried out for the MiCId workflow. We thank the staff of the Culture Collection University of Gothenburg (CCUG, Gothenburg, Sweden) for providing bacterial strains. The CCUG is supported by the Department of Clinical Microbiology, Sahlgrenska University Hospital and the Sahlgrenska Academy of the University of Gothenburg, Sweden. R.K., D.J.L., B.P.I., F.S.S., and E.R.B.M. acknowledge support and funding from the Center for Antibiotic Resistance Research (CARe, Sahlgrenska Academy, University of Gothenburg). We thank the Proteomics Core Facility at the Sahlgrenska Academy, University of Gothenburg, for performing proteomics experiments and proteomics analysis using the Proteome Discoverer software version 2.4 (Thermo Fisher Scientific). This work was supported by the Intramural Research Program of the National Library of Medicine. Funding for Open Access publication charges for this article was provided by the National Institutes of Health.

■ REFERENCES

(1) French, G. L. Clinical impact and relevance of antibiotic resistance. *Adv. Drug Deliv Rev.* **2005**, *57* (10), 1514−1527.

(2) Cosgrove, S. E. The relationship between antimicrobial resistance and patient outcomes: mortality, length of hospital stay, and health care costs. *Clin Infect Dis* **2006**, *42* (Suppl 2), S82−89.

(3) Lode, H. M. Clinical impact of antibiotic-resistant Gram-positive pathogens. *Clin Microbiol Infect* **2009**, *15* (3), 212−217.

(4) Ciorba, V.; Odone, A.; Veronesi, L.; Pasquarella, C.; Signorelli, C. Antibiotic resistance as a major public health concern: epidemiology and economic impact. *Ann. Ig* **2015**, *27* (3), 562−579.

(5) Schneider, J. E.; Romanowsky, J.; Schuetz, P.; Stojanovic, I.; Cheng, H. K.; Liesenfeld, O.; Buturovic, L.; Sweeney, T. E. Cost Impact Model of a Novel Multi-mRNA Host Response Assay for Diagnosis and Risk Assessment of Acute Respiratory Tract Infections and Sepsis in the Emergency Department. *J. Health Econ Outcomes Res.* **2020**, *7* (1), 24−34.

(6) Isakov, O.; Modai, S.; Shomron, N. Pathogen detection using short-RNA deep sequencing subtraction and assembly. *Bioinformatics* **2011**, *27* (15), 2027−2030.

(7) Ammerlaan, H. S. M.; Harbarth, S.; Buiting, A. G. M.; Crook, D. W.; Fitzpatrick, F.; Hanberger, H.; Herwaldt, L. A.; van Keulen, P. H. J.; Kluytmans, J. A. J. W.; Kola, A.; Kuchenbecker, R. S.; Lingaas, E.; Meessen, N.; Morris-Downes, M. M.; Pottinger, J. M.; Rohner, P.; dos Santos, R. P.; Seifert, H.; Wisplinghoff, H.; Ziesing, S.; Walker, A. S.; Bonten, M. J. M. Secular Trends in Nosocomial Bloodstream Infections: Antibiotic-Resistant Bacteria Increase the Total Burden of Infection. *Clinical Infectious Diseases* **2013**, *56* (6), 798−805.

(8) Dworzanski, J. P.; Deshpande, S. V.; Chen, R.; Jabbour, R. E.; Snyder, A. P.; Wick, C. H.; Li, L. Mass spectrometry-based proteomics combined with bioinformatic tools for bacterial classification. *J. Proteome Res.* **2006**, *5* (1), 76−87.

(9) Sauer, S.; Kliem, M. Mass spectrometry tools for the classification and identification of bacteria. *Nat. Rev. Microbiol.* **2010**, *8* (1), 74−82.

(10) Giebel, R., Worden, C., Rust, S., Kleinheinz, G., Robbins, M., Sandrin, T. Microbial Fingerprinting Using Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS): Applications and Challenges. *Advances in Applied Microbiology*; Academic Press, 2010; Vol. *71*, pp 149 − 184.

(11) Bazinet, A. L.; Cummings, M. P. A comparative evaluation of sequence classification programs. *BMC Bioinformatics* **2012**, *13*, 92.

(12) Miller, R. R.; Montoya, V.; Gardy, J. L.; Patrick, D. M.; Tang, P. Metagenomics for pathogen detection in public health. *Genome Med.* **2013**, *5* (9), 81.

(13) Penzlin, A.; Lindner, M. S.; Doellinger, J.; Dabrowski, P. W.; Nitsche, A.; Renard, B. Y. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. *Bioinformatics* **2014**, *30* (12), i149−156.

(14) Naccache, S. N.; Federman, S.; Veeraraghavan, N.; Zaharia, M.; Lee, D.; Samayoa, E.; Bouquet, J.; Greninger, A. L.; Luk, K. C.; Enge, B.; Wadford, D. A.; Messenger, S. L.; Genrich, G. L.; Pellegrino, K.; Grard, G.; Leroy, E.; Schneider, B. S.; Fair, J. N.; Martinez, M. A.; Isa, P.; Crump, J. A.; DeRisi, J. L.; Sittler, T.; Hackett, J.; Miller, S.; Chiu, C. Y. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* **2014**, *24* (7), 1180−1192.

(15) Dudhagara, P.; Bhavsar, S.; Bhagat, C.; Ghelani, A.; Bhatt, S.; Patel, R. Web Resources for Metagenomics Studies. *Genomics, Proteomics & Bioinformatics* **2015**, *13* (5), 296−303.

(16) Karlsson, R.; Gonzales-Siles, L.; Boulund, F.; Svensson-Stadler, L.; Skovbjerg, S.; Karlsson, A.; Davidson, M.; Hulth, S.; Kristiansson, E.; Moore, E. R. Proteotyping: Proteomic characterization, classification and identification of microorganisms−A prospectus. *Syst. Appl. Microbiol.* **2015**, *38* (4), 246−257.

(17) Srinivasan, R.; Karaoz, U.; Volegova, M.; MacKichan, J.; Kato-Maeda, M.; Miller, S.; Nadarajan, R.; Brodie, E. L.; Lynch, S. V. Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. *PLoS One* **2015**, *10* (2), No. e0117617.

(18) Singhal, N.; Kumar, M.; Kanaujia, P. K.; Virdi, J. S. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front Microbiol* **2015**, *6*, 791.

(19) Opota, O.; Jaton, K.; Greub, G. Microbial diagnosis of bloodstream infection: towards molecular diagnosis directly from blood. *Clin. Microbiol. Infect.* **2015**, *21* (4), 323−331.

(20) Mesuere, B.; Debyser, G.; Aerts, M.; Devreese, B.; Vandamme, P.; Dawyndt, P. The Unipept metaproteomics analysis pipeline. *Proteomics* **2015**, *15* (8), 1437−1442.

(21) Alves, G.; Wang, G.; Ogurtsov, A. Y.; Drake, S. K.; Gucek, M.; Suffredini, A. F.; Sacks, D. B.; Yu, Y. K. Identification of Microorganisms by High Resolution Tandem Mass Spectrometry with

Accurate Statistical Significance. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (2), 194−210.
(22) Zhang, X.; Ning, Z.; Mayne, J.; Moore, J. I.; Li, J.; Butcher, J.; Deeke, S. A.; Chen, R.; Chiang, C. K.; Wen, M.; Mack, D.; Stintzi, A.; Figeys, D. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **2016**, *4* (1), 31.
(23) Anjum, M. F.; Zankari, E.; Hasman, H. Molecular Methods for Detection of Antimicrobial Resistance. *Microbiol Spectr* **2017**, DOI: 10.1128/microbiolspec.ARBA-0011-2017.
(24) Boulund, F.; Karlsson, R.; Gonzales-Siles, L.; Johnning, A.; Karami, N.; Al-Bayati, O.; Ahren, C.; Moore, E. R. B.; Kristiansson, E. TCUP: Typing and characterization of bacteria using bottom-up tandem mass spectrometry proteomics. *Mol. Cell Proteomics* **2017**, *16*, 1052.
(25) Wolff, N.; Hendling, M.; Schroeder, F.; Schönthaler, S.; Geiss, A. F.; Bedenic, B.; Barišic, I. Full pathogen characterisation: species identification including the detection of virulence factors and antibiotic resistance genes via multiplex DNA-assays. *Sci. Rep* **2021**, *11* (1), 6001.
(26) Jagtap, P. D.; Blakely, A.; Murray, K.; Stewart, S.; Kooren, J.; Johnson, J. E.; Rhodus, N. L.; Rudney, J.; Griffin, T. J. Metaproteomic analysis using the Galaxy framework. *Proteomics* **2015**, *15* (20), 3553−3565.
(27) Bortolaia, V.; Kaas, R. S.; Ruppe, E.; Roberts, M. C.; Schwarz, S.; Cattoir, V.; Philippon, A.; Allesoe, R. L.; Rebelo, A. R.; Florensa, A. F.; Fagelhauer, L.; Chakraborty, T.; Neumann, B.; Werner, G.; Bender, J. K.; Stingl, K.; Nguyen, M.; Coppens, J.; Xavier, B. B.; Malhotra-Kumar, S.; Westh, H.; Pinholt, M.; Anjum, M. F.; Duggett, N. A.; Kempf, I.; Nykäsenoja, S.; Olkkola, S.; Wieczorek, K.; Amaro, A.; Clemente, L.; Mossong, J.; Losch, S.; Ragimbeau, C.; Lund, O.; Aarestrup, F. M. ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **2020**, *75* (12), 3491−3500.
(28) Niu, S. Y.; Yang, J.; McDermaid, A.; Zhao, J.; Kang, Y.; Ma, Q. Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Brief Bioinform* **2018**, *19* (6), 1415−1429.
(29) Alves, G.; Wang, G.; Ogurtsov, A. Y.; Drake, S. K.; Gucek, M.; Sacks, D. B.; Yu, Y. K. Rapid Classification and Identification of Multiple Microorganisms with Accurate Statistical Significance via High-Resolution Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (8), 1721−135.
(30) Alves, G.; Yu, Y. K. Robust Accurate Identification and Biomass Estimates of Microorganisms via Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (1), 85−102.
(31) Suh, M. J.; Keasey, S. L.; Brueggemann, E. E.; Ulrich, R. G. Antibiotic-dependent perturbations of extended spectrum beta-lactamase producing Klebsiella pneumoniae proteome. *Proteomics* **2017**, *17* (9), 1700003.
(32) Keasey, S. L.; Suh, M. J.; Das, S.; Blancett, C. D.; Zeng, X.; Andresson, T.; Sun, M. G.; Ulrich, R. G. Decreased Antibiotic Susceptibility Driven by Global Remodeling of the Klebsiella pneumoniae Proteome. *Mol. Cell Proteomics* **2019**, *18* (4), 657−668.
(33) Chen, C. Y.; Clark, C. G.; Langner, S.; Boyd, D. A.; Bharat, A.; McCorrister, S. J.; McArthur, A. G.; Graham, M. R.; Westmacott, G. R.; Van Domselaar, G. Detection of Antimicrobial Resistance Using Proteomics and the Comprehensive Antibiotic Resistance Database: A Case Study. *Proteomics Clin Appl.* **2020**, *14* (4), No. e1800182.
(34) Yu, Y.; O'Rourke, A.; Lin, Y. H.; Singh, H.; Eguez, R. V.; Beyhan, S.; Nelson, K. E. Predictive Signatures of 19 Antibiotic-Induced Escherichia coli Proteomes. *ACS Infect Dis* **2020**, *6* (8), 2120−2129.
(35) Kulkarni, P.; Frommolt, P. Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. *Comput. Struct Biotechnol J.* **2017**, *15*, 471−477.
(36) Heyer, R.; Schallert, K.; Zoun, R.; Becher, B.; Saake, G.; Benndorf, D. Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **2017**, *261*, 24−36.
(37) Schiebenhoefer, H.; Van Den Bossche, T.; Fuchs, S.; Renard, B. Y.; Muth, T.; Martens, L. Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Rev. Proteomics* **2019**, *16* (5), 375−390.
(38) Chatterjee, S.; Stupp, G. S.; Park, S. K.; Ducom, J. C.; Yates, J. R.; Su, A. I.; Wolan, D. W. A comprehensive and scalable database search system for metaproteomics. *BMC Genomics* **2016**, *17* (1), 642.
(39) Kleiner, M.; Thorson, E.; Sharp, C. E.; Dong, X.; Liu, D.; Li, C.; Strous, M. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **2017**, *8* (1), 1558.
(40) Alves, G.; Yu, Y. K. Mass spectrometry-based protein identification with accurate statistical significance assignment. *Bioinformatics* **2015**, *31* (5), 699−706.
(41) Zhang, G.; Ueberheide, B. M.; Waldemarson, S.; Myung, S.; Molloy, K.; Eriksson, J.; Chait, B. T.; Neubert, T. A.; Fenyo, D. Methods Mol BiolProtein quantitation using mass spectrometry. *Methods Mol. Biol.* **2010**, *673*, 211−222.
(42) van de Merbel, N. C. BioanalysisProtein quantification by LC-MS: a decade of progress through the pages of Bioanalysis. *Bioanalysis* **2019**, *11* (7), 629−644.
(43) McArdle, A. J.; Turkova, A.; Cunnington, A. J. When do co-infections matter? *Curr. Opin Infect Dis* **2018**, *31* (3), 209−215.
(44) Handel, A.; Longini, I. M.; Antia, R. Intervention strategies for an influenza pandemic taking into account secondary bacterial infections. *Epidemics* **2009**, *1* (3), 185−195.
(45) Manohar, P.; Loh, B.; Athira, S.; Nachimuthu, R.; Hua, X.; Welburn, S. C.; Leptihn, S. Secondary Bacterial Infections During Pulmonary Viral Disease: Phage Therapeutics as Alternatives to Antibiotics? *Front Microbiol* **2020**, *11*, 1434.
(46) Wilmes, P.; Bond, P. L. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol* **2006**, *14* (2), 92−97.
(47) Salvato, F.; Hettich, R. L.; Kleiner, M. Five key aspects of metaproteomics as a tool to understand functional interactions in host-associated microbiomes. *PLoS Pathog* **2021**, *17* (2), No. e1009245.
(48) Kondori, N.; Kurtovic, A.; Pieiro-Iglesias, B.; Salv-Serra, F.; Jaén-Luchoro, D.; Andersson, B.; Alves, G.; Ogurtsov, A.; Thorsell, A.; Fuchs, J.; Tunovic, T.; Kamenska, N.; Karlsson, A.; Yu, Y. K.; Moore, E.R. B.; Karlsson, R. in Blood. *Front Cell Infect Microbiol* **2021**, *11*, 634215.
(49) Alves, G.; Ogurtsov, A. Y.; Yu, Y. K. RAId_DbS: peptide identification using database searches with realistic statistics. *Biol. Direct* **2007**, *2*, 25.
(50) Ogunseitan, O. A. Bacterial genetic exchange in nature. *Sci. Prog.* **1995**, *78* (Part 3), 183−204.
(51) Thomas, C. M.; Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol* **2005**, *3* (9), 711−721.
(52) Alcock, B. P.; Raphenya, A. R.; Lau, T.T. Y.; Tsang, K. K.; Bouchard, M.; Edalatmand, A.; Huynh, W.; Nguyen, A. V.; Cheng, A. A.; Liu, S.; Min, S. Y.; Miroshnichenko, A.; Tran, H. K.; Werfalli, R. E.; Nasir, J. A.; Oloni, M.; Speicher, D. J.; Florescu, A.; Singh, B.; Faltyn, M.; Hernandez-Koutoucheva, A.; Sharma, A. N.; Bordeleau, E.; Pawlowski, A. C.; Zubyk, H. L.; Dooley, D.; Griffiths, E.; Maguire, F.; Winsor, G. L.; Beiko, R. G.; Brinkman, F.S. L.; Hsiao, W.W. L.; Domselaar, G. V.; McArthur, A. G. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **2020**, *48* (D1), D517−D525.
(53) Sayers, E. W.; Beck, J.; Brister, J. R.; Bolton, E. E.; Canese, K.; Comeau, D. C.; Funk, K.; Ketter, A.; Kim, S.; Kimchi, A.; Kitts, P. A.; Kuznetsov, A.; Lathrop, S.; Lu, Z.; McGarvey, K.; Madden, T. L.; Murphy, T. D.; OLeary, N.; Phan, L.; Schneider, V. A.; Thibaud-Nissen, F.; Trawick, B. W.; Pruitt, K. D.; Ostell, J. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2019**, *48* (D1), D9−D16.
(54) Prokaryotic RefSeq Genomes. https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/.
(55) Lukjancenko, O.; Wassenaar, T. M.; Ussery, D. W. Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol* **2010**, *60* (4), 708−720.

(56) Kaas, R. S.; Friis, C.; Ussery, D. W.; Aarestrup, F. M. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. *BMC Genomics* **2012**, *13*, 577.
(57) Bush, K.; Jacoby, G. A. Updated functional classification of beta-lactamases. *Antimicrob. Agents Chemother.* **2010**, *54* (3), 969−976.
(58) Huang, T.; Wang, J.; Yu, W.; He, Z. Protein inference: a review. *Brief Bioinform* **2012**, *13* (5), 586−614.
(59) Vizcano, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ros, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P. A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H. J.; Albar, J. P.; Martinez-Bartolomé, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223−226.
(60) Giske, C. G.; Rylander, M.; Kronvall, G. VIM-4 in a carbapenem-resistant strain of Pseudomonas aeruginosa isolated in Sweden. *Antimicrob. Agents Chemother.* **2003**, *47* (9), 3034−3035.
(61) Johnning, A.; Jakobsson, H. E.; Boulund, F.; Salvà-Serra, F.; Moore, E. R.; Åhrén, C.; Karami, N.; Kristiansson, E. Draft Genome Sequence of Extended-Spectrum-β-Lactamase-Producing Escherichia coli Strain CCUG 62462, Isolated from a Urine Sample. *Genome Announc* **2016**, *4* (6), 223.
(62) Johnning, A.; Karami, N.; Tång Hallbäck, E.; Müller, V.; Nyberg, L.; Buongermino Pereira, M.; Stewart, C.; Ambjörnsson, T.; Westerlund, F.; Adlerberth, I.; Kristiansson, E. The resistomes of six carbapenem-resistant pathogens - a critical genotype-phenotype analysis. *Microb Genom* **2018**, *4* (11), e000233.
(63) O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; Astashyn, A.; Badretdin, A.; Bao, Y.; Blinkova, O.; Brover, V.; Chetvernin, V.; Choi, J.; Cox, E.; Ermolaeva, O.; Farrell, C. M.; Goldfarb, T.; Gupta, T.; Haft, D.; Hatcher, E.; Hlavina, W.; Joardar, V. S.; Kodali, V. K.; Li, W.; Maglott, D.; Masterson, P.; McGarvey, K. M.; Murphy, M. R.; O'Neill, K.; Pujar, S.; Rangwala, S. H.; Rausch, D.; Riddick, L. D.; Schoch, C.; Shkeda, A.; Storz, S. S.; Sun, H.; Thibaud-Nissen, F.; Tolstoy, I.; Tully, R. E.; Vatsan, A. R.; Wallin, C.; Webb, D.; Wu, W.; Landrum, M. J.; Kimchi, A.; Tatusova, T.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44* (D1), D733−745.
(64) Girlich, D.; Naas, T.; Nordmann, P. Biochemical characterization of the naturally occurring oxacillinase OXA-50 of Pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* **2004**, *48* (6), 2043−2048.
(65) Rodríguez-Martnez, J. M.; Poirel, L.; Nordmann, P. Extended-spectrum cephalosporinases in Pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* **2009**, *53* (5), 1766−1771.
(66) Sentausa, E., Basso, P., Berry, A., Adrait, A., Bellement, G., Couté, Y., Lory, S., Elsen, S., Attrée, I.: Insertion sequences drive the emergence of a highly adapted human pathogen. *Microb Genom 6* (9) (**2020**).mgen000265
(67) National Database of Antibiotic Resistant Organisms (NDARO). https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/.
(68) Lan, R.; Reeves, P. R. Escherichia coli in disguise: molecular origins of Shigella. *Microbes Infect.* **2002**, *4* (11), 1125−1132.
(69) WHO publishes list of bacteria for which new antibiotics are urgently needed. https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed.
(70) Mingeot-Leclercq, M. P.; Glupczynski, Y.; Tulkens, P. M. Aminoglycosides: activity and resistance. *Antimicrob. Agents Chemother.* **1999**, *43* (4), 727−737.
(71) Ramirez, M. S.; Tolmasky, M. E. Aminoglycoside modifying enzymes. *Drug Resist Updat* **2010**, *13* (6), 151−171.
(72) Karlsson, R.; Gonzales-Siles, L.; Gomila, M.; Busquets, A.; Salva-Serra, F.; Jaen-Luchoro, D.; Jakobsson, H. E.; Karlsson, A.; Boulund, F.; Kristiansson, E.; Moore, E.R. B. Proteotyping bacteria: Characterization, differentiation and identification of pneumococcus and other species within the Mitis Group of the genus Streptococcus by tandem mass spectrometry proteomics. *PLoS One* **2018**, *13* (12), No. e0208804.
(73) Eliuk, S.; Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal Chem. (Palo Alto Calif)* **2015**, *8*, 61−80.
(74) Chenau, J.; Fenaille, F.; Caro, V.; Haustant, M.; Diancourt, L.; Klee, S. R.; Junot, C.; Ezan, E.; Goossens, P. L.; Becher, F. Identification and validation of specific markers of Bacillus anthracis spores by proteomics and genomics approaches. *Mol. Cell Proteomics* **2014**, *13* (3), 716−732.
(75) Chen, S. H.; Parker, C. H.; Croley, T. R.; McFarland, M. A. Identification of Salmonella Taxon-Specific Peptide Markers to the Serovar Level by Mass Spectrometry. *Anal. Chem.* **2019**, *91* (7), 4388−4395.
(76) Karlsson, R.; Thorsell, A.; Gomila, M.; Salvà-Serra, F.; Jakobsson, H. E.; Gonzales-Siles, L.; Jaén-Luchoro, D.; Skovbjerg, S.; Fuchs, J.; Karlsson, A.; Boulund, F.; Johnning, A.; Kristiansson, E.; Moore, E.R. B. Discovery of Species-unique Peptide Biomarkers of Bacterial Pathogens by Tandem Mass Spectrometry-based Proteotyping. *Mol. Cell Proteomics* **2020**, *19* (3), 518−528.
(77) Grossegesse, M.; Hartkopf, F.; Nitsche, A.; Schaade, L.; Doellinger, J.; Muth, T. Perspective on Proteomics for Virus Detection in Clinical Samples. *J. Proteome Res.* **2020**, *19* (11), 4380−4388.
(78) Stackebrandt, E.; Goebel, B. M. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology* **1994**, *44* (4), 846−849.
(79) Bull, M. J.; Marchesi, J. R.; Vandamme, P.; Plummer, S.; Mahenthiralingam, E. Minimum taxonomic criteria for bacterial genome sequence depositions and announcements. *J. Microbiol Methods* **2012**, *89* (1), 18−21.
(80) Schoch, C. L.; Ciufo, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; Sharma, S.; Soussov, V.; Sullivan, J. P.; Sun, L.; Turner, S.; Karsch-Mizrachi, I. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* **2020**, *2020*, baaa062.
(81) Barco, R. A.; Garrity, G. M.; Scott, J. J.; Amend, J. P.; Nealson, K. H.; Emerson, D. A Genus Definition for Bacteria and Archaea Based on a Standard Genome Relatedness Index. *mBio* **2020**, DOI: 10.1128/mBio.02475-19.
(82) Pible, O.; Hartmann, E. M.; Imbert, G.; Armengaud, J. The importance of recognizing and reporting sequence database contamination for proteomics. *EuPA Open Proteomics* **2014**, *3*, 246−249.
(83) Schäffer, A. A.; Nawrocki, E. P.; Choi, Y.; Kitts, P. A.; Karsch-Mizrachi, I.; McVeigh, R. VecScreen_plus_taxonomy: imposing a tax(onomy) increase on vector contamination screening. *Bioinformatics* **2018**, *34* (5), 755−759.
(84) Alves, G.; Ogurtsov, A. Y.; Yu, Y. K. RAId_DbS: mass-spectrometry based peptide identification web server with knowledge integration. *BMC Genomics* **2008**, *9*, 505.
(85) Sorić, B. Statistical "Discoveries" and Effect-Size Estimation. *Journal of the American Statistical Association* **1989**, *84* (406), 608−610.
(86) Tharakan, R.; Edwards, N.; Graham, D. R. Data maximization by multipass analysis of protein mass spectra. *Proteomics* **2010**, *10* (6), 1160−1171.
(87) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (7), 1111−1120.
(88) Bern, M.; Kil, Y. J. Comment on "Unbiased statistical analysis for multi-stage proteomic search strategies. *J. Proteome Res.* **2011**, *10* (4), 2123−2127.
(89) Alves, G.; Yu, Y.-K. Improving Peptide Identification Sensitivity in Shotgun Proteomics by Stratification of Search Space. *J. Proteome Res.* **2013**, *12* (6), 2571−2581.
(90) Ghafourian, S.; Sadeghifard, N.; Soheili, S.; Sekawi, Z. Extended Spectrum Beta-lactamases: Definition, Classification and Epidemiology. *Curr. Issues Mol. Biol.* **2015**, *17*, 11−21.
(91) Bush, K.; Bradford, P. A. β-Lactams and β-Lactamase Inhibitors: An Overview. *Cold Spring Harb Perspect Med.* **2016**, *6* (8), a025247.

(92) Papp-Wallace, K. M.; Endimiani, A.; Taracila, M. A.; Bonomo, R. A. Carbapenems: past, present, and future. *Antimicrob. Agents Chemother.* **2011**, *55* (11), 4943−4960.
(93) Bonnet, R. Growing group of extended-spectrum beta-lactamases: the CTX-M enzymes. *Antimicrob. Agents Chemother.* **2004**, *48* (1), 1−14.
(94) Paterson, D. L.; Bonomo, R. A. Extended-spectrum β-lactamases: a clinical update. *Clin Microbiol Rev.* **2005**, *18* (4), 657−686.
(95) Kong, K. F.; Jayawardena, S. R.; Del Puerto, A.; Wiehlmann, L.; Laabs, U.; Tmmler, B.; Mathee, K. Characterization of poxB, a chromosomal-encoded Pseudomonas aeruginosa oxacillinase. *Gene* **2005**, *358*, 82−92.
(96) Rather, P. N.; Parojcic, M. M.; Paradise, M. R. An extracellular factor regulating expression of the chromosomal aminoglycoside 2'-N-acetyltransferase of Providencia stuartii. *Antimicrob. Agents Chemother.* **1997**, *41* (8), 1749−1754.
(97) Wright, G. D.; Ladak, P. Overexpression and characterization of the chromosomal aminoglycoside 6'-N-acetyltransferase from Enterococcus faecium. *Antimicrob. Agents Chemother.* **1997**, *41* (5), 956−960.
(98) Zincke, D.; Balasubramanian, D.; Silver, L. L.; Mathee, K. Characterization of a Carbapenem-Hydrolyzing Enzyme, PoxB, in Pseudomonas aeruginosa PAO1. *Antimicrob. Agents Chemother.* **2016**, *60* (2), 936−945.
(99) Blanco, P.; Hernando-Amado, S.; Reales-Calderon, J. A.; Corona, F.; Lira, F.; Alcalde-Rico, M.; Bernardini, A.; Sanchez, M. B.; Martinez, J. L. Bacterial Multidrug Efflux Pumps: Much More Than Antibiotic Resistance Determinants. *Microorganisms* **2016**, *4* (1), 14.
(100) Sun, Q.; Ba, Z.; Wu, G.; Wang, W.; Lin, S.; Yang, H. Insertion sequence ISRP10 inactivation of the oprD gene in imipenem-resistant Pseudomonas aeruginosa clinical isolates. *Int. J. Antimicrob. Agents* **2016**, *47* (5), 375−379.
(101) Wolkowicz, T.; Patzer, J. A.; Kaminska, W.; Gierczynski, R.; Dzierzanowska, D. Distribution of carbapenem resistance mechanisms in Pseudomonas aeruginosa isolates among hospitalised children in Poland: Characterisation of two novel insertion sequences disrupting the oprD gene. *J. Glob Antimicrob Resist* **2016**, *7*, 119−125.
(102) Lindberg, F.; Lindquist, S.; Normark, S. Inactivation of the ampD gene causes semiconstitutive overproduction of the inducible Citrobacter freundii beta-lactamase. *J. Bacteriol.* **1987**, *169* (5), 1923−1928.
(103) Pérez-Gallego, M.; Torrens, G.; Castillo-Vera, J.; Moya, B.; Zamorano, L.; Cabot, G.; Hultenby, K.; Albertí, S.; Mellroth, P.; Henriques-Normark, B.; Normark, S.; Oliver, A.; Juan, C. Impact of AmpC Derepression on Fitness and Virulence: the Mechanism or the Pathway? *mBio* **2016**, *7* (5), e01783-16 DOI: 10.1128/mBio.01783-16.
(104) Casabona, M. G.; Vandenbrouck, Y.; Attree, I.; Couté, Y. Proteomic characterization of Pseudomonas aeruginosa PAO1 inner membrane. *Proteomics* **2013**, *13* (16), 2419−2423.
(105) Malherbe, G.; Humphreys, D. P.; Davé, E. A robust fractionation method for protein subcellular localization studies in Escherichia coli. *Biotechniques* **2019**, *66* (4), 171−178.
(106) Thein, M.; Sauer, G.; Paramasivam, N.; Grin, I.; Linke, D. Efficient subfractionation of gram-negative bacteria for proteomics studies. *J. Proteome Res.* **2010**, *9* (12), 6135−6147.
(107) Kulak, N. A.; Geyer, P. E.; Mann, M. Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics. *Mol. Cell Proteomics* **2017**, *16* (4), 694−705.

https://doi.org/10.1021/jasms.1c00347
J. Am. Soc. Mass Spectrom. 2022, 33, 917−931